Tutorial

Analysis of SARS-CoV-2 using MinION sequences

December 18, 2023

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Analysis of SARS-CoV-2 using MinION sequences

Sequencing using MinION is a fast and cost-efficient way to track the evolution of SARS-CoV-2 as it allows sequencing to be done in a few hours with little investment in laboratory equipment. However, samples are often of low quality and contaminated with RNA from other sources. Here we demonstrate how to overcome the challenge of metatranscriptomics and get the most from Oxford Nanopore MinION reads using the *CLC Genomics Workbench*.

The tutorial covers the following:

- Import of data required for the analysis.
- Trimming MinION reads.
- Mapping MinION reads to a reference.
- Calling variants in the sample relative to a reference and visualizing variant calls and mappings.
- Extracting a consensus sequence from a read mapping.
- Using BLAST to identify a strain.
- Running the pipeline in a workflow.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 24.0 or higher with the Long Read Support plugin installed.

How to install plugins using the Plugin Manager is described here: https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html

Plugins can also be downloaded from (https://digitalinsights.qiagen.com/products-overview/plugins/) and then installed using the Plugin Manager.

If you are not already familiar with the tools in the Long Read Support plugin, we suggest first completing the tutorial "De Novo Assembly Using Long Reads and Short Read Polishing".

Background information

The data you will be working with is from metatranscriptome sequencing using a MinION data set from the Wuhan seafood market pneumonia outbreak [Chan et al., 2020]. Sampling the virus responsible for the disease involves obtaining sputum, throat or nasopharyngeal swabs, or bronchoalveolar lavage fluid (BALF) samples from a patient. Because of this, samples will contain RNA from non-viral sources. This data set was prepared using the Sequence-Independent, Single-Primer Amplification (SISPA) protocol for additional viral sequence enrichment.

Some basic tips

- In this tutorial, we refer to the tools in the Workbench menus, but you can instead click on the **Launch** button (🚀) in the top toolbar to search for and launch tools of interest.

- We will run several of the tools in "Batch" mode. Using the "Batch" option allows us to analyze several input elements individually, with the same parameter settings, while launching the tool wizard only once.

Import the data

1. Download the example data from our web site https://resources.qiagenbioinformatics.com/testdata/SARS-CoV-2_MinION_example_data.zip and unzip it.

2. Open the *CLC Genomics Workbench*.

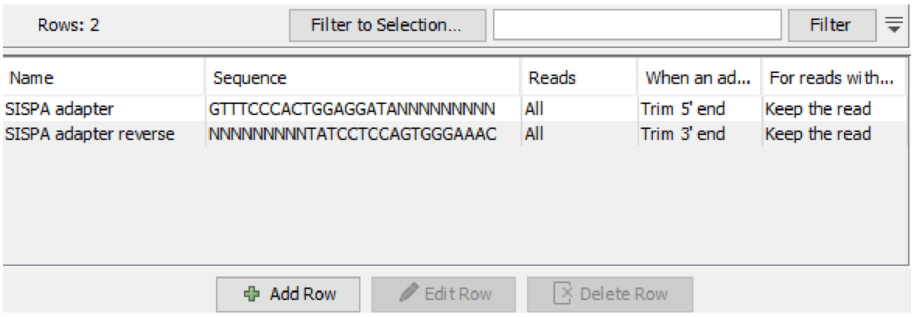
3. Create a new folder for the project with a relevant name, for example named "SARS-CoV-2 MinION Tutorial".

4. Import the reference sequence and adapter list provided with the example data by going to:

   File | **Import** | **Standard Import**

   Choose the option **Automatic import** and select "MT135044.gbk" and "SISPA adapter trim list.clc". Save the items in the folder you just created.

   After import, you can open the SISPA adapter trim list and see that it looks like the one shown in figure 1.



Name	Sequence	Reads	When an ad...	For reads with...
SISPA adapter	GTTTCCCACTGGAGGATANNNNNNNNN	All	Trim 5' end	Keep the read
SISPA adapter reverse	NNNNNNNNNTATCCTCCAGTGGGAAAC	All	Trim 3' end	Keep the read

Figure 1: *Adapter trim list.*

5. Download the full dataset containing the MinION reads and sample metadata using SRA.

   - Search for reads in SRA by going to:

     Download | **Search for Reads in SRA** (**)**

   - Choose the term "Bioproject" in the drop-down menu in the top left corner and enter "PRJNA601630" into the search field.

   - In the results table, select the rows with Run accessions "SRR10948474" and "SRR10948550" and click on **Download Reads and Metadata** in the bottom of the view. Leave the options set as the defaults. Save the results to a new folder, for example named "Raw Reads".

   If, for some reason, you are unable to access SRA we have included the raw reads in the example data.

   - To use the raw reads from the example data, go to:

     File | **Import** | **Oxford Nanopore**

- Select "SRR10948550.fastq" and "SRR10948474.fastq". Save to a new folder, for example named "Raw Reads".

Information on these read sets

Run	# of Spots	# of Bases	Size	Sampling method
SRR10948474	505,484	284.6M	250.1Mb	Sputum
SRR10948550	425,717	146.3M	126.6Mb	Nasopharyngeal swab

Trim adapters from the reads

The samples were prepared using a SISPA protocol, so the first thing we will do is trim the reads using the Trim Reads tool and the imported Trim Adapter List.

1. Run the Trim Reads tool to trim adapters from both sets of reads.
   - Go to:

     Toolbox | **Prepare Sequencing Data ()** | **Trim Reads ()**
   - Select the sequence lists containing the reads and move these to the *Selected elements* field on the right. Click on **Next**.
   - Deselect both of the options in the *Quality trimming* step and then click on **Next**.
   - In the *Adapter trimming* step of the wizard:
     - Click on the file selector icon () to the right of the **Trim adapter list** field then select the SISPA adapter trim list and click on **OK**.
     - Click on **Next**.
   - Leave options in the *Trim homopolymers* step and *Sequence filtering* steps unchecked and click on **Next** in each case.
   - In the Results handling steps, select the **Create report** option and choose to save the output to a new folder, for example named "Trimmed reads".

Map long reads to reference

In this section, you will map the reads to a closely related strain to create a new consensus sequence. We will use isolate "MT135044" as the reference. This was imported in an earlier step. We stress that this is an example, and any high-quality assembly of a closely related strain will work.
Note: Since we are dealing with metatranscriptomic data, we do not expect all reads to map and coverage may be quite low as a result.

1. Run the Map Long Reads to Reference tool by going to:

   Toolbox | **Long Read Support ()** | **Map Long Reads to Reference ()**

2. Select the trimmed reads created in the previous step. Check **Batch** to run once per sample and click **Next**.

3. Leave the *Batch overview* settings as is and click **Next**.

4. In *References*, use the file selector icon () to select the previously imported MT135044 reference sequence and click **OK** then **Next**.

5. Leave the mapping options set to **Automatic** and click **Next**.

6. Check the **Create report** option and choose to save the results in a specified folder. Click **Next** and save the results to a new folder, for example named "Read mappings". You can monitor the progress of the read mapping job under the *Processes* tab in the bottom left of the Workbench. The tool will take approximately 1-2 minutes per sample depending on your hardware.

7. After the tool has finished running, inspect the mapping reports. The table below shows an example of the read mapping statistics:

   Read mapping report

Run	Count	% reads	Average Length	# of bases	% bases
SRR10948474	191,114	37.91%	663.11	126,730,061	62.34%
SRR10948550	996	0.23%	702.25	699,445	0.80%

It can be seen that a large percentage of reads in SRR10948550 are not from SARS-CoV-2.

Call variants relative to the reference sequence

1. Next, you will call variants on these read mappings using the Fixed Ploidy Variant Detection tool:

   Resequencing Analysis () | **Variant Detection ()** | **Fixed Ploidy Variant Detection ()**

2. Select the read mappings and choose to run the job in batch mode.

3. Change the *Fixed ploidy variant parameters* to **Ploidy** = 1 and **Required variant probability (%)** = 80. Click **Next**.

4. Under *Coverage and count filters* set **Minimum count** = 7 and **Minimum frequency (%)** = 75.0. The other options can be left as default. Click **Next**.

5. Disable all options in *Noise filters* and click **Next**.

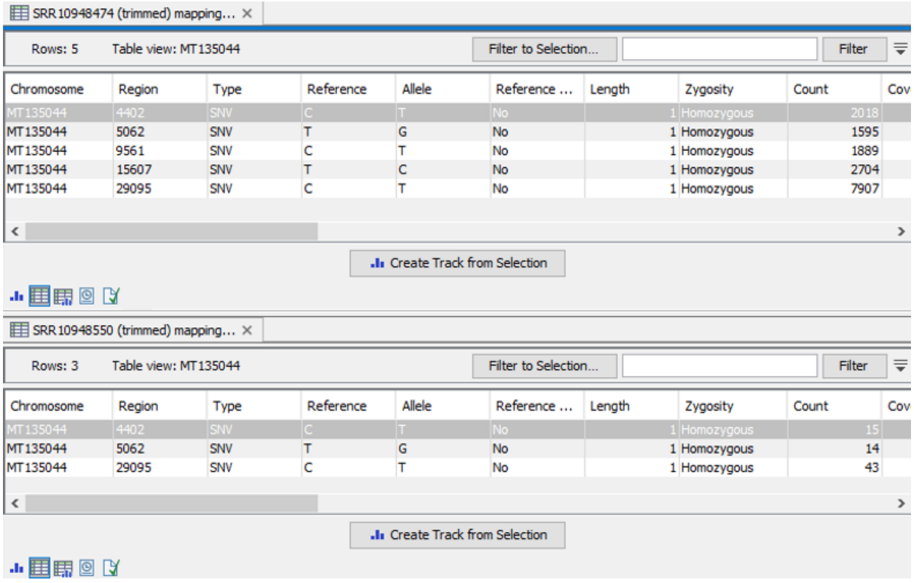
6. Make sure the **Create track** option is selected and choose to save the variant calls in a new folder, for example named "SARS-CoV-2 variant tracks".

7. Open the output variant tracks to observe the result. Both runs have variant calls relative to the MT135044 reference, three of which are present in both samples. Figure 2 shows the variant tracks in Table View for both samples.

   The sampling dates are January 28, 2020 for MT135044 and before January 21, 2020 for SRR10948474 and SRR10948550. This shows the ability to trace mutations over time during an active outbreak.

8. Now the results can be inspected together by selecting the read mappings () and variant tracks () for both samples and executing **File** | **New** | **Track List ()**. Keep the track list open for now. You can also choose to save it by pressing CTRL + S (Cmd + S on Mac).

9. Optional: You can include amino acid changes by first creating a genome and CDS track.



SRR10948474 (trimmed) mapping...

Rows: 5 Table view: MT135044

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Cov
MT135044	4402	SNV	C	T	No	1	Homozygous	2018	
MT135044	5062	SNV	T	G	No	1	Homozygous	1595	
MT135044	9561	SNV	C	T	No	1	Homozygous	1889	
MT135044	15607	SNV	T	C	No	1	Homozygous	2704	
MT135044	29095	SNV	C	T	No	1	Homozygous	7907	

SRR10948550 (trimmed) mapping...

Rows: 3 Table view: MT135044

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Cov
MT135044	4402	SNV	C	T	No	1	Homozygous	15	
MT135044	5062	SNV	T	G	No	1	Homozygous	14	
MT135044	29095	SNV	C	T	No	1	Homozygous	43	

Figure 2: *Table view of the variant tracks. Top: SRR10948474, Bottom: SRR10948550.*

- To start, open:

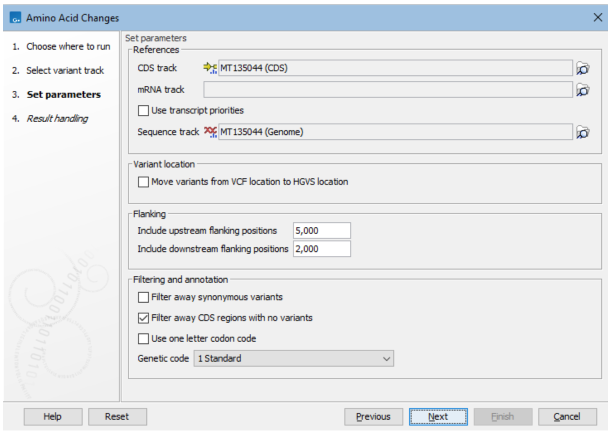
  Utility Tools () | **Tracks** () | **Track Conversion** () | **Convert to Tracks** ()
- Select MT135044 as input and click **Next**.
- Check **Create sequence track** and **Create annotation tracks**. In **Annotation types**, press (✚), select CDS and click **Done** to close the dialog. Then click **Next**.
- Choose to save the result, click **Next** and save to a new folder, for example named "Tracks".
- Next, annotate variants and create a track for visual inspection by running:

  Resequencing Analysis () | **Functional Consequences** () | **Amino Acid Changes** ()
- Remember to run in batch mode and select the variant tracks. Click **Next**.
- In *Set parameters* select the CDS and Genome tracks created in previous steps as shown on figure 3. Leave the other settings on default. Click **Next**, then save to your variant location.
- You can now drag and drop the amino acids tracks into the open track list.

10. The track list view can be used to investigate the read mappings, the read coverage across a variant position, the effect on amino acid sequence due to variants, etc. If you zoom in on the variant in position 9561, your track list should look like figure 4. For details on how to work with tracks, see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Tracks.html`.
    Tip: As MinION reads have many errors, a more stringent display of the read mapping gives a better experience. To achieve this, change the settings in the **Side Panel** to the right. Under *Track layout* | *Reads track* | *Hide insertions below (%)* change the value to e.g., 80.

Extract consensus sequences from the read mappings

We have shown how to use an existing reference to call variants. This section focuses on how to use the read mapping to construct consensus sequences for these samples. The consensus



Amino Acid Changes

1. Choose where to run
2. Select variant track
3. **Set parameters**
4. *Result handling*

Set parameters

References

CDS track: MT135044 (CDS)

mRNA track:

□ Use transcript priorities

Sequence track: MT135044 (Genome)

Variant location

□ Move variants from VCF location to HGVS location

Flanking

Include upstream flanking positions: 5,000

Include downstream flanking positions: 2,000

Filtering and annotation

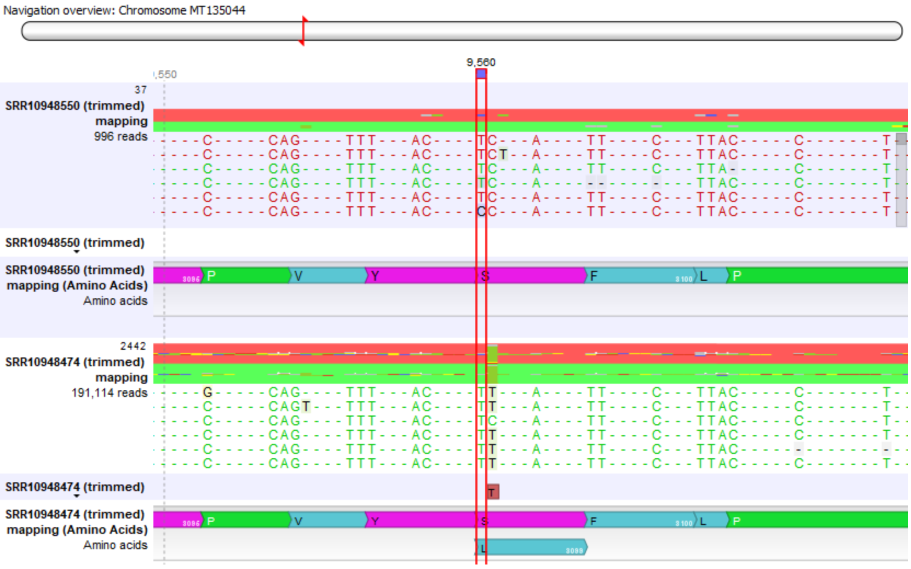
□ Filter away synonymous variants

☑ Filter away CDS regions with no variants

□ Use one letter codon code

Genetic code: 1 Standard

Figure 3: *Amino Acid Changes options.*



Navigation overview: Chromosome MT135044
9,560
9,550
37
SRR10948550 (trimmed) mapping
996 reads
SRR10948550 (trimmed)
SRR10948550 (trimmed) mapping (Amino Acids)
Amino acids
2442
SRR10948474 (trimmed) mapping
191,114 reads
SRR10948474 (trimmed)
SRR10948474 (trimmed) mapping (Amino Acids)
Amino acids

Figure 4: *Read mapping of MinION samples and variant call visualization.*

sequences will be used in the next steps to search for similar sequences in a custom BLAST database.

1. To begin, go to:

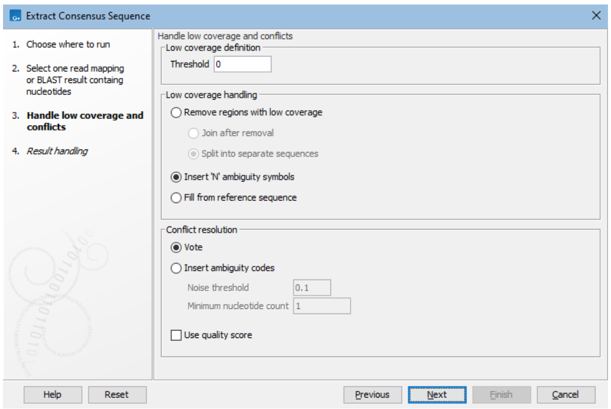
   Resequencing Analysis ( ) | Extract Consensus Sequence ( )

   Check **Batch** to run once per sample and select the read mappings created in a previous step. Click **Next**.

2. Leave the options in *Handle low coverage and conflicts* set to their default values. Your options screen should look like figure 5. Click **Next**.

3. Select to save the consensus sequences in a new folder, for example named "Consensus sequences".

You should now have two complete, high quality sequences.



1. Choose where to run
2. Select one read mapping or BLAST result containg nucleotides
3. **Handle low coverage and conflicts**
4. *Result handling*

Handle low coverage and conflicts

Low coverage definition

Threshold 0

Low coverage handling

○ Remove regions with low coverage

○ Join after removal

● Split into separate sequences

● Insert 'N' ambiguity symbols

○ Fill from reference sequence

Conflict resolution

● Vote

○ Insert ambiguity codes

Noise threshold 0.1

Minimum nucleotide count 1

□ Use quality score

Figure 5: *Extract Consensus Sequence options.*

Using BLAST to compare to other SARS-CoV-2 assemblies

The sequences can be used as queries in a BLAST search to see which assemblies they match. For this tutorial, we will run a local BLAST search against a custom BLAST database, which we will create first. It is also possible to download one of NCBI's public databases e.g., Betacoronavirus, for this step, but given the large size of the database and the length of the sequences, the download and search will take several hours.

For this example, we will use the sequence list "Custom Betacoronavirus Database" from the tutorial zip file, which contains a list of Genbank assemblies with the taxonomical genus level "Betacoronavirus".

1. Locate the sequence list "Custom Betacoronavirus Database", which you imported in the beginning of the tutorial, and use it as input for the Create BLAST Database tool:

   Toolbox | **BLAST ()**| **Create BLAST Database ()**

   Click **Next**. If you want, you can change the name and description, but it is not necessary. Then click **Finish**.

2. Now, run BLAST on the two consensus sequences we generated in the previous step. Launch BLAST by going to:

   Toolbox | **BLAST ()**| **BLAST ()**

3. Choose both of the consensus sequence lists as the input for the tool and click on **Next**.

4. Choose "blastn" as the type of *BLAST program* to run and select the option **BLAST database** as the *Target* type. From the drop-down menu, select the custom BLAST database you just created, then click on **Next**.

5. We are looking for sequences similar to our query sequence and are only interested in the top hits, so change **Expect** = 0.01 and **Max number of hit sequences** = 100. Leave the other search settings as default, and click on **Next**.

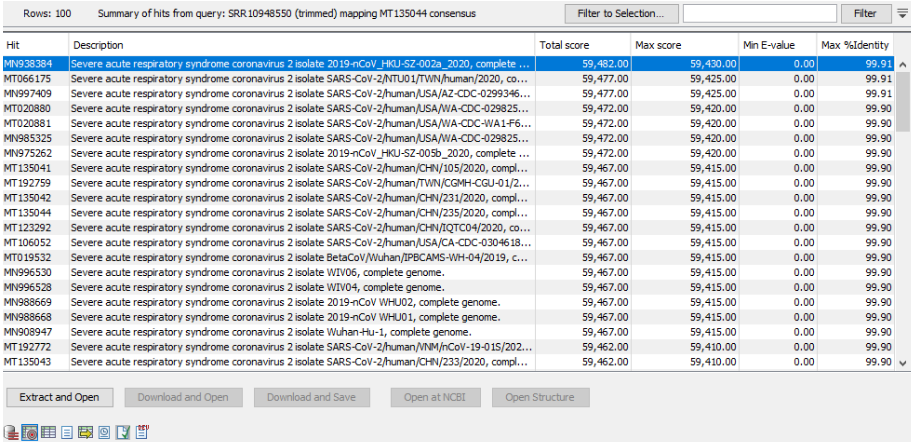
6. Choose to **Open** the results rather than saving them, and click on **Finish**.

7. A Multi BLAST Table containing the two queries and the top reported result will open. Take note of the best matching sequence in the *Accession (E-value)* column. For SRR10948550,

we expect to see MN938384, and for SRR10948474, we expect to see MN975262. These are the sequences submitted by [Chan et al., 2020] that were assembled based on the same read sets we are using.

More information about the best matches can be seen in this table by ticking the boxes in the *Show column* section in the **Side Panel**.

8. Press the **Open BLAST Output** buttom in the bottom of the Multi BLAST Table view. This will open an individual table for the selected query (figure 6). In the bottom left of the new table, you can choose between three different views, *BLAST Graphics*, *BLAST Hit Table*, and *BLAST HSP Table*, which hold analysis values such as Max score, %Identity, %Positive, %Gaps, etc.



Rows: 100 Summary of hits from query: SRR10948550 (trimmed) mapping MT135044 consensus

Hit	Description	Total score	Max score	Min E-value	Max %Identity
MN938384	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-002a_2020, complete ...	59,482.00	59,430.00	0.00	99.91
MT066175	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/NTU01/TWN/human/2020, co...	59,477.00	59,425.00	0.00	99.91
MN997409	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/AZ-CDC-0299346...	59,477.00	59,425.00	0.00	99.91
MT020880	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-029825...	59,472.00	59,420.00	0.00	99.90
MT020881	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-WA1-F6...	59,472.00	59,420.00	0.00	99.90
MN985325	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-029825...	59,472.00	59,420.00	0.00	99.90
MN975262	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-005b_2020, complete ...	59,472.00	59,420.00	0.00	99.90
MT135041	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/105/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT192759	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/TWN/CGMH-CGU-01/2...	59,467.00	59,415.00	0.00	99.90
MT135042	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/231/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT135044	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/235/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT123292	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/IQTC04/2020, co...	59,467.00	59,415.00	0.00	99.90
MT106052	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-0304618...	59,467.00	59,415.00	0.00	99.90
MT019532	Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/IPBCAMS-WH-04/2019, c...	59,467.00	59,415.00	0.00	99.90
MN996530	Severe acute respiratory syndrome coronavirus 2 isolate WIV06, complete genome.	59,467.00	59,415.00	0.00	99.90
MN996528	Severe acute respiratory syndrome coronavirus 2 isolate WIV04, complete genome.	59,467.00	59,415.00	0.00	99.90
MN988669	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU02, complete genome.	59,467.00	59,415.00	0.00	99.90
MN988668	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01, complete genome.	59,467.00	59,415.00	0.00	99.90
MN908947	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.	59,467.00	59,415.00	0.00	99.90
MT192772	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/VNM/nCoV-19-01S/202...	59,462.00	59,410.00	0.00	99.90
MT135043	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/233/2020, compl...	59,462.00	59,410.00	0.00	99.90

Figure 6: *BLAST output hit table.*

9. You can retreive the sequence of any of the BLAST hits by selecting the hit and clicking on the **Extract and Open** button. You could choose to save the best hit for each sample for further analysis.

You have now identified the known strains most closely related to your samples.

Run the analysis using a workflow

To automate the process for reproducible execution of the trimming, mapping, variant calling, and consensus sequence extraction, we have provided a workflow in the zip file. The steps to download a BLAST database and run a BLAST search have to be executed manually.
In this workflow we also added QC tools, which can be used to evaluate results. For information on these tools, see the *CLC Genomics Workbench* user manual:

- **QC for Sequencing Reads**: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Sequencing_Reads.html
- **QC for Read Mapping**: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html

Workflows can be installed using the Workflow manager, see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Installing_workflow.html.

Installed workflows are available from the **Toolbox** under **Installed workflows**. You may need to update the workflow after installing.

You can inspect the workflow by double-clicking on it, and you can run it on any number of samples in batch mode. On the MinION reads provided in the introduction, this workflow takes less than 10 minutes on a 2013 MacBook pro laptop computer (2,6 GHz Intel Core i5).

Figure 7 shows the automated pipeline of the bioinformatics steps in the workflow editor.

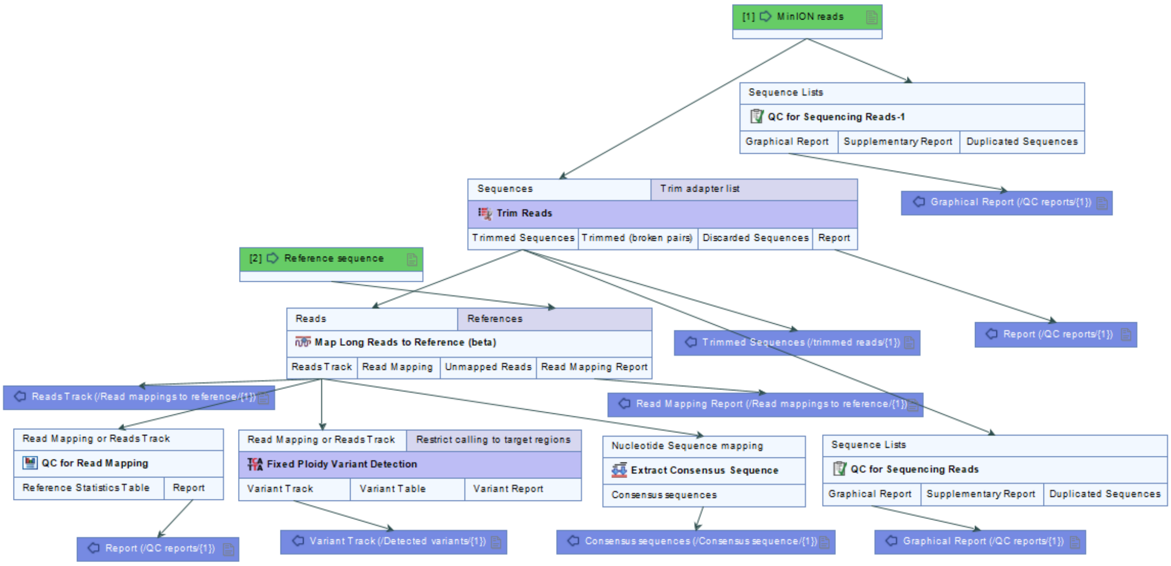

[1] MinION reads
Sequence Lists
QC for Sequencing Reads-1
Graphical Report
Supplementary Report
Duplicated Sequences
Sequences
Trim adapter list
Trim Reads
Trimmed Sequences
Trimmed (broken pairs)
Discarded Sequences
Report
Graphical Report (/QC reports/{1})
[2] Reference sequence
Reads
References
Map Long Reads to Reference (beta)
Reads Track
Read Mapping
Unmapped Reads
Read Mapping Report
Trimmed Sequences (/trimmed reads/{1})
Report (/QC reports/{1})
Reads Track (/Read mappings to reference/{1})
Read Mapping Report (/Read mappings to reference/{1})
Read Mapping or Reads Track
QC for Read Mapping
Reference Statistics Table
Report
Read Mapping or Reads Track
Restrict calling to target regions
Fixed Ploidy Variant Detection
Variant Track
Variant Table
Variant Report
Nucleotide Sequence mapping
Extract Consensus Sequence
Consensus sequences
Sequence Lists
QC for Sequencing Reads
Graphical Report
Supplementary Report
Duplicated Sequences
Report (/QC reports/{1})
Variant Track (/Detected variants/{1})
Consensus sequences (/Consensus sequence/{1})
Graphical Report (/QC reports/{1})

Figure 7: *A workflow reproducing the bioinformatics pipeline.*

Bibliography

[Chan et al., 2020] Chan, J., Yuan, S., Kok, K., To, K., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C., Poon, R., Tsoi, H., Lo, S., Chan, K., Poon, V., Chan, W., Ip, J., Cai, J., Cheng, V., Chen, H., Hui, C., and Yuen, K. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, 395(15):514–523.